Building a COVID-19 Web Archive with Grant Funding

Lyndsey Blair, Web Archiving Assistant
Claire Drone-Silvers, Web Archiving Assistant
Denise Rayman, Digital Archivist and Philanthropic Studies Librarian
Rhys Weber, Web Archiving Assistant

Project Background

Shortly after the World Health Organization (WHO) declared COVID-19 a pandemic in March 2020, the archival community set off to preserve information about this event for the historic record. Many of the COVID-19 collections developed in the aftermath of this announcement have been local in nature, focusing on a specific community, city, or state's response to the pandemic. In Indianapolis, the Ruth Lilly Special Collections & Archives at IUPUI University Library has spent the last year building its own web archive about the COVID-19 response in Central Indiana through a health, public health, social, and economic lens.

This project is generously funded by the Central Indiana Community Foundation Library Fund (Central Indiana Community Foundation, n.d.) for the labor and data, and a COVID-19 collection cost-sharing program from Archive-It (Archive-It Staff, 2020).

The project is available to view at https://archive-it.org/collections/14457

About Archive-It

Archive-It is a service of the Internet Archive, home of the Wayback Machine. While not the only web archiving option, they are the largest. Internet Archive is also a nonprofit with a Platinum transparency rating (Candid, 2020) from Guidestar (Candid, n.d.), a nonprofit reporting database.

Subscription model, but no ongoing cost commitment

Archive-It works on a yearly subscription model, but unlike your other subscription costs with vendors, they never delete data. (Archive-It Staff, 2021) For example, after talking to staff, you estimate you will crawl 1TB of data for your project and sign up for a subscription for the year at that level. At the end of your project, all the data you have captured is saved forever. If you end your subscription or drop to a lower data level, all the material you saved in the last cycle has no ongoing cost to you.

Suitable for grant projects

Explaining archival work to grantors can often be difficult, but the high public visibility of the Internet Archive's work (especially the Wayback Machine) is worthy of note. While explaining the financial attractiveness of a comprehensive digital preservation plan is difficult, most of the lay public have experienced a dead URL, or read about major web deletions like Yahoo Answers (Garcia, 2021). The pay-once-saved-forever subscription model makes Archive-It a good partner for grant projects, as ongoing financial support of previous grant projects is a problem for many archives. The Internet Archive's status as a non-profit can also help when working with grantors.

Flexible and easy to use

The Internet Archives provides a wealth of information to new users to help get started, and in-depth troubleshooting assistance while using the service, including email support tickets as well as Zoom "Office Hour" appointments. None of our team members had recent web archiving experience prior to starting the

project, so the instructional videos and resource pages from the Help Center were exceedingly beneficial. The Archive-It service also offers flexibility to its partners by allowing users to implement their own metadata standards, collection development policies, and collecting strategies and to select specific information for the web crawler to capture.

Core tasks in developing a new web archive collection

Collection Development

Selection of what material to capture (and not capture) is the first step in any web archiving project, as well as an ongoing decision. As will be expounded later, setting up both functional and ethical guidelines to seed selection is a must for any web archiving team. Any project is going to have limits on both data and labor, meaning that collection development will need to be respectful of both the financial limits of a project as well as the number of people involved. In ethical terms, the data we archive should be ethically selected and gathered, as well as respectful of individuals' needs and privacy.

Seed scoping, design, and testing

Assigning the seed type and adding scope rules to a particular seed allows us to determine the direction and size of each captured crawl. For example, a seed set to the Standard+ configuration will capture more information in a crawl than a seed set to the One Page configuration; therefore, you might add specific scoping rules to the Standard+ seed in order to capture specific documents or to refrain from capturing excessive or unnecessary information (such as links to social media or video sites). If you are particularly concerned about a crawl consuming too much data, setting crawl limits (such as the length of time the crawler can run, or a data cap) can provide security against an unintentional data binge.

We cannot overstate the importance of running a one-time test crawl prior to permanently adding a seed to the collection or changing its seed type. Test crawls are designed to be deleted by default; that way, if a test crawl goes wild or does not properly capture the seed, you can make changes to the seed type or seed scoping rules and run another test crawl without the risk of depleting your data allotment. If a test crawl runs smoothly, however, you can choose to save the entire test crawl to your collections, and then decide how frequently to crawl it after the initial capture (one-time, weekly, monthly).

Quality assurance

Quality assurance (QA) is the area where a new web archiving project is most likely to run into trouble. While seed scope design and testing will hopefully sort out any serious issues, we have found that even minor shifts in website layout, design, and content can have drastic consequences. Therefore, the focus of QA work must be split between both fixing crawl issues with missing images and other documents, alongside making sure that unnecessary or irrelevant data is *not* crawled, which wastes data. Quality assurance overlaps with the previous task of seed scope design and testing. While scope settings are meant to be set up prior to a seed's first crawl, often a website will change significantly enough to require additional QA work in adjusting the seed scope rules as a website evolves.

Metadata creation

The Archive-It metadata interface is simple to use and is customizable. Collection development and QA have been our biggest priority for the past few months, but we have also collaborated with other IUPUI units to determine how to format the collection-level and seed-level metadata. We selected the appropriate metadata fields based on the University Library's metadata policy and Archive-It's standard format, which

both use Dublin Core. While specific metadata fields can be made private, Titles and Descriptions can be browsed by the public and should provide clear and practical information to users. Based on our project's needs, we decided to add optional metadata fields, such as "Crawling Technology Used," in order to provide helpful information for users that could explain why one monthly crawl looks different from the next one. Our current metadata fields for individual seeds are:

- Collector (our archives)
- Crawling Technology Used
- Creator (LCNAF)
- Date
- Description
- Identifier (URL)
- Language
- Subject (LCSH, when possible)
- Title
- Notes

Estimating labor and data costs for web archiving

Writing an accurate and workable budget is the foundation of a viable grant project, but nearly impossible to do for an entirely new venture. As we had never done web archiving before, our original budget has been modified several times. While the varying nature of web material means web archiving costs do not scale predictably either by number of URL seeds or amount of data, we can pass on some wisdom to help others estimate costs for a web archiving project.

This grant officially began on April 1, 2020, when the full length of the COVID-19 pandemic was not known. As such our data budget has been worked backwards: we have a set amount of data we can capture, and we had to appraise material for collection based within that limit. We have since added more data budget from our original April 2020 request, but this is obviously not an ideal way to do collection development. If you have time, it is possible to do a data estimate on websites before formally crawling them, using the Test Crawl feature in Archive-It, running a locally-stored test crawl on your own machine with Heritrix (Osborne, 2021), or even using the ArchiveWeb.page browser extension to create a single WARC file (Webrecorder.net, n.d.). These methods can be used to create a data estimate before writing a grant budget.

The largest investment in this project has far and away been labor. The Web Archiving Assistants, working part time on this project, have collectively logged over 600 hours, and are currently working about 60 hours per week. The costliest labor investment is quality assurance (QA) of crawled material. We have about 300 seeds, 160 of which are ongoing (weekly or monthly) periodical crawls, and the others are one-time crawls. A reoccurring crawl is more expensive both in labor and data, and so when designing a web archiving project, reoccurring crawls should be selected very carefully. We estimate that it can take 30-60 minutes to set up and test an initial URL seed, assuming it crawls well and there are no errors. For a single troublesome website, it can take several hours to troubleshoot and patch-repair a bad crawl.

Appraisal and collection of web material

We made our selection of URL seeds and material to crawl for this collection with great care. As this is an ongoing event of mass death and trauma, our first consideration was to collect material ethically. Fortunately, many archivists had been working and publishing on ethical contemporaneous collecting around

the Black Lives Matter movement before the COVID-19 pandemic, and we were able to use these emerging ethical guidelines to inform our appraisal decisions. The work of Project STAND (Project STAND (Student Activism Now Documented), 2018) and DocNow (Jules et al., 2018) are especially helpful.

Due to concerns about personal privacy, we decided not to target any individual experiences for collection, such as personal blogs, personal social media accounts, or oral histories. While this material will be critical for archives to collect, it will need to be done slowly, with thoughtfulness and informed consent from individuals, not as part of massive web crawls (Tansey, 2020). State and local governments, businesses, nonprofit organizations, school districts, and large religious bodies were the target of our crawling campaign.

We were also thoughtful of the traditional appraisal concern about duplication of information. We did not crawl any newspaper sites, as from our experience we expect that material to be widely collected by others, both in print and digital. We also did not target government publications and datasets, for the same reason that they will be collected by other parties. Due to the price of data, we also limited our collecting of video content to things we thought were exceptionally unique and likely to be valued by researchers.

Our collecting focused on what we thought was going to disappear and change quickly and not be captured through any other means. Some of the material we captured has already disappeared or heavily changed. We are particularly proud of our collection of school websites, which captured the chaotic response to COVID-19 from the elementary to college level, across public and private schools, and several Indiana counties.

Lessons and Takeaways

Technical Challenges

The technical issue that we struggled with most during this project was unintentionally large crawls. Even if a crawl was set up carefully and tested before being put into production, if they website changed too much between the original set up and time of crawl, the software could accidentally crawl enormous amounts of irrelevant or frivolous data. For our team's favorite example, when executing a routine crawl of a page about Coronavirus disaster recovery, the crawler somehow captured a video titled "12 Hours of Cricket Sound Effects in Stereo". While humorous in retrospect, at the time it was maddening to try and find just how the crawler had managed to find such a large file, especially one that had no affiliation with what that seed had been intended to crawl. What we learned from crawls like this, was that even crawls that seem safe and contained often have the capacity to go off the rails.

Accepting the "Good Enough" Crawl

After spending hours troubleshooting and re-crawling sites, we often had to accept that Archive-It as a crawling technology was not always able to properly capture certain sites. For example, we have had quite a few seeds where the text of a crawled seed appeared fine, but the general layout of the page or images had issues. Oftentimes if the text or charts were readable, the general layout or images were put aside, especially if the data had been crawled but was just not properly displayed. This was an area where we had to decide what a "good enough" crawl was. We eventually decided that we should focus on the data and information that future researchers might find useful, and that capturing the exact look and feel of websites was ideal, but not likely to be as important to researchers as actual content.

Undone Work

Writing both collection and seed level metadata is where we have the most work yet to do. We decided it is better to do work slowly than to have to redo metadata down the line, so we are working on developing our

metadata fields and standards before writing any actual metadata. Our library has an established system and process for metadata creation with new digital collections, however, that team had never worked with web material before, and wanted time to decide on metadata standards.

Due to the lack of metadata, our collection, while technically public, is nearly impossible to browse and use, so we haven't been advertising it widely until it can be properly described. However, since our metadata standards will be applied to any new web archiving work, our next large web archiving venture will be able to have metadata created as it grows, making it usable from the start.

Bibliography

Archive-It Staff. (2020, March). COVID-19 Web Archiving Special Campaign. *Archive-It Blog*. https://archive-it.org/blog/covid-19/

Archive-It Staff. (2021, March 31). *Want to know more about Archive-It? Subscription FAQ*. Archive-It Help Center. https://support.archive-it.org/hc/en-us/articles/208111766-Want-to-know-more-about-Archive-It-

Candid. (n.d.). *INTERNET ARCHIVE - Profile*. Guidestar. Retrieved April 23, 2021, from https://www.guidestar.org/profile/94-3242767

Candid. (2020, March). *About the GuideStar Seals of Transparency*. https://learn.guidestar.org/seals

Central Indiana Community Foundation. (n.d.). *About The Indianapolis Foundation Library Fund*. Central Indiana Community Foundation. Retrieved April 23, 2021, from https://www.cicf.org/about-cicf/funds-and-foundations/special-focus-funds/the-indianapolis-foundation-library-fund/

Garcia, R. T. (2021, April 20). *Deleting Yahoo Answers is a disastrous idea. For history's sake, we need to preserve our digital record.* Business Insider. https://www.businessinsider.com/deleting-yahoo-answers-disastrous-idea-preserve-our-digital-record-2021-4

Jules, B., Summers, E., & Mitchell Jr., V. (2018). *Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations* [White Paper]. Documenting the Now (DocNow). https://www.docnow.io/docs/docnow-whitepaper-2018.pdf

Osborne, A. (2021, February 15). *About Heritrix*. GitHub. https://github.com/internetarchive/heritrix3

Project STAND (Student Activism Now Documented). (2018, August 11). *S.A.V.E METHODOLOGY*. Project STAND. https://standarchives.com/s-a-v-e-methodology/

Tansey, E. (2020, June 5). No one owes their trauma to archivists, or, the commodification of contemporaneous collecting. *Eira Tansey*. https://eiratansey.com/2020/06/05/no-one-owes-their-trauma-to-archivists-or-the-commodification-of-contemporaneous-collecting/

Webrecorder.net. (n.d.). *ArchiveWeb.page Extension: User Guide*. Retrieved April 27, 2021, from https://archiveweb.page/guide/